Counterfactual Reasoning for Responsible AI Assessment

Giandomenico Cornacchia[1], Vito Walter Anelli[1], Fedelucio Narducci[1], Azzurra Ragone[2] and Eugenio Di Sciascio[1]

[1]Polytechnic University of Bari, Via Orabona, 4, Bari, 70125, Italy
[2]Università degli Studi di Bari Aldo Moro, Piazza Umberto I, 1, Bari, 70125, Italy

Abstract
As the use of AI and ML models continues to grow, concerns about potential unfairness have become more prominent. Many researchers have focused on developing new definitions of fairness or identifying biased predictions, but these approaches have limited scope and fail to analyze the minimum changes in user characteristics required for positive outcomes (i.e. counterfactuals). In response, this proposed methodology aims to use counterfactual reasoning to identify unfair behaviours in the case of fairness under unawareness. Furthermore, counterfactual reasoning can serve as a comprehensive methodology for evaluating all the essential conditions for a reliable, responsible, and trustworthy model.

Keywords
Counterfactual Reasoning, Fairness, Audit, Explainability, Responsibility

1. Introduction

As stated by the World Economic Forum's Global Future Council on Artificial Intelligence for Humanity"*Artificial Intelligence (AI) is the engine of the Fourth Industrial Revolution. It holds the promise of solving some of society's most pressing issues, including repowering economies reeling from lockdowns, but requires thoughtful design, development, and deployment to mitigate potential risks*"[1].

These risks are related to the fact that AI applications are becoming more and more pervasive, and, most of the time, users often interact with such systems without even knowing that life-changing decisions like mortgage grants, job offers, patients screenings are in the hand of AI-based systems. Moreover, such AI decisions may sometimes result arbitrary, inconsistent, or discriminatory, which cannot be allowed in highly regulated environments such as Financial Services. As these applications have became key enablers and more deeply embedded in processes, financial services organizations need to cope with AI applications' inherent risks. This is true both from a compliance point of view (regulatory and ethical norms), and because the lack of trust is the most significant barrier to AI adoption and acceptance by users. In fact, AI systems often amplify social and ethical issues such as gender and demographic discrimination, and they lack interpretability and explainability.

As an example, in the financial domain, the decision to approve or deny credit has been regulated with precise and detailed regulatory compliance requirements (i.e., Equal Credit Opportunity Act , Federal Fair Lending Act , and Consumer Credit Directive for EU Community).These rules aim to prevent discrimination in human decision-making processes. However, they do not fit scenarios involving Machine Learning (ML) or, more broadly, Artificial Intelligence (AI) systems. However, when AI replaces human decisions, like in the case of instant lending, there is a risk of revealing a loophole in existing liability identification laws. Several national and international organizations have released guidelines, norms, and principles to prevent the irresponsible usage of AI, e.g., the EU Commission with "The Proposal for Harmonized Rule on AI" and the expert group on "AI in Society" of the Organisation for Economic Co-operation and Development (OECD).

Although scientists train their models without explicit discriminating intent, deploying AI systems without taking ethical concerns into account may lead to discrimination [1]. Even more problematic is figuring out which type of discrimination is being implemented.

1.1. Counterfactual Reasoning as a Responsible AI practice

Counterfactual Reasoning is an active and flourishing field in artificial intelligence research [2, 3]. This research was initially born to investigate causal links [4], and today it can count on several contributions [5]. Most of them define and employ counterfactuals as a helpful

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy
giandomenico.cornacchia@poliba.it (G. Cornacchia); vitowalter.anelli@poliba.it (V. W. Anelli); fedelucio.narducci@poliba.it (F. Narducci); azzurra.ragone@uniba.it (A. Ragone); eugenio.disciascio@poliba.it (E. D. Sciascio)
0000-0001-5448-9970 (G. Cornacchia); 0000-0002-5567-4307 (V. W. Anelli); 0000-0002-9255-3256 (F. Narducci); 0000-0002-3537-7663 (A. Ragone); 0000-0002-5484-9945 (E. D. Sciascio)

[1]https://www.weforum.org/communities/gfc-on-artificial-intelligence-for-humanity

tools to explain the decisions taken by modern decision support systems. The underlying rationale is that some aspects of past events could predict future events. In detail, some studies focus on identifying causality-related aspects to discover the link between the counterfactuals and the analyzed phenomenon.

Counterfactual Reasoning finds application in various fields. To summarize what we have briefly detailed before, machine learning research has positively valued these contributions ranging from Explainable AI [6] to the most recent counterfactual fairness measures [7, 8].

Beyond the theoretical aspects, Counterfactual Reasoning is extensively applied to interactive systems [9, 10, 11, 12]. Unfortunately, this important application showed some limitations. These systems employ machine learning models that reflect the data they use for learning. Consequently, the same information influences the reasoning, and the contribution of Counterfactual Reasoning could be limited or somehow biased. The explaining policy, coming from Counterfactual Reasoning, exhibits a bias toward the implemented learning model. Researchers devoted considerable effort to tackle this issue and proposed new models such as doubly robust estimators [13]. Overall, even though limitations that need a solution, Counterfactual Reasoning is taking over Explainable AI, and it is becoming the de facto standard for explaining decisions taken by autonomous systems [14]. In this respect, the European Union's "right to explanation" played a crucial role in arousing a further interest in this methodologies [15]. Indeed, they are compliant with the regulation and easily interpreted by either a domain expert or a layperson [16].

Decision support systems particularly benefited from these models. However, the more the application domain is vital, the more the fairness problem emerges. For instance, the issue cannot be overlooked in sensitive domains such as justice, risk assessment, or clinical risk prediction. This need promoted the most promising research in the Counterfactual Reasoning field to analyze and mitigate this issue. A further important issue under the lens of European regulators is the discrimination of AI models. On this point, the EU Commission proposes a conformity assessment before AI systems are put into service or placed on the market [2]. In fact, their tools are subject to fair and trustworthy audit assessments to check their conformity. However, is a shallow check of the input characteristics sufficient to determine that a predictor will not suggest unfair treatment? Even though the user does not provide protected characteristics, the system could predict sensitive features from variables, i.e., proxy variables, that still represent protected characteristics[17, 18, 19]. In this regard, our investigation aims to leverage a counterfactual generation tool to reveal the presence of implicit biases in a decision support system. The approach aims to answer the question: "How would the system have decided if we had replaced some user characteristics? These characteristics identify a protected or a non-protected group?".

[2]https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

2. Preliminaries

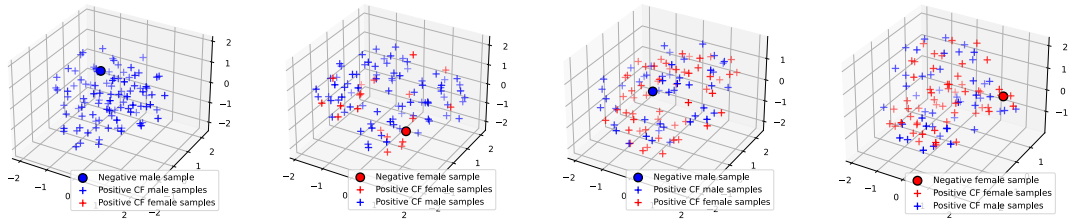
This section introduces the notation adopted hereinafter.
Data points: We assume the dataset $\mathscr{D}$ is an m-dimensional space containing n non-sensitive features, l sensitive features, and a target attribute. In other words, we have $\mathscr{D} \subseteq \mathbb{R}^m$, with $m = n+l+1$. A data point $d \in \mathscr{D}$ is then represented as $d = \langle \mathbf{x}, \mathbf{s}, y \rangle$, with $\mathbf{x} = \langle x_1, x_2, ..., x_n \rangle$ representing the sub-vector of non-sensitive features, $\mathbf{s} = \langle s_1, s_2, ..., s_l \rangle$ the sub-vector of sensitive features and y being a binary target feature. Given a vector of sensitive features, $\forall s_i \in \mathbf{s}$, $s_i = 0$ refers to the *unprivileged* group and $s_i = 1$ to the *privileged* group of the i-th sensitive feature.
Target Labels: Given a target feature $y \in \{0, 1\}$, $y = 1$ is the positive outcome and $y = 0$ is the negative one.
Outcome Prediction: $\hat{y} \in \{0, 1\}$ represents the prediction for $\mathbf{x} \subset d$ estimated by $f(\cdot)$, a function such that $f(\mathbf{x}) = \hat{y}$.
Sensitive Feature Prediction: $\hat{s}_i \in \{0, 1\}$ represents the prediction of the i-th sensitive feature for a given data point estimated by $f_{s_i}(\cdot)$, a function s.t. $f_{s_i}(\mathbf{x}) = \hat{s}_i$.
Counterfactual samples: Given a vector $\mathbf{x}$ and a perturbation $\epsilon = \langle \epsilon_1, \epsilon_2, ..., \epsilon_n \rangle$, we say that a vector $\mathbf{c_x} = \langle c_{x_1}, c_{x_2}, ..., c_{x_n} \rangle = \mathbf{x} + \epsilon$ is a counterfactual (CF) of $\mathbf{x}$ if $f(\mathbf{c_x}) = 1 - f(\mathbf{x}) = 1 - \hat{y}$. We use the set $\mathscr{C}_\mathbf{x}$, with $|\mathscr{C}_\mathbf{x}| = k$, to denote the set of possible **counterfactual samples** for $\mathbf{x}$. A function $g(\mathbf{x})$ compute k counterfactuals for $\mathbf{x}$.
For simplicity, we denote $f(\cdot)$, $f_{s_i}(\cdot)$, and $g(\cdot)$ as the **Decision Maker**, the **Sensitive-Feature Classifier**, and the **Counterfactual Generator** respectively.

3. Methodology

Our study proposes a novel fairness definition, two novel metrics for detecting bias in a scenario where sensitive features are omitted (i.e., *fairness under unawareness*) in the training process, and an explanation methodology.

3.1. Fairness through the counterfactual lens

Excluding sensitive features makes verifying that all users are treated equally incredibly challenging. In the instant lending case, imagine that a customer applies for a loan, and his/her request is rejected. Understanding



(a) male on Classic ML model (b) female on Classic ML model (c) male on Debiasing model (d) female on Debiasing model

Figure 1: Adult t-SNE visualizations of a random male (a-c) and female (b-d) sample with a negative outcome and their CF samples with a positive outcome, respectively, for a Classic ML model (i.e. XGB) and a Debiasing one (i.e. Adversarial Debiasing).

if the customer has been discriminated is hard to verify when sensitive information is not used. Our process pipeline is as follows: the **Decision Maker** makes decisions without exploiting sensitive features, then if the outcome is negative (e.g. loan rejected), the **Counterfactual Generator** is exploited to propose modifications to user characteristics and request for reaching a positive outcome (e.g. loan approved). For each data point d with a negative prediction $f(\mathbf{x}) = 0$, we generate a set of counterfactual samples $\mathscr{C}_\mathbf{x}$ that reach a positive outcome (i.e., $\forall \mathbf{c}_\mathbf{x} \in \mathscr{C}_\mathbf{x}$ s.t. $f(\mathbf{c}_\mathbf{x}) = 1$). Afterward, each counterfactual (CF) sample is evaluated by the **Sensitive-Feature Classifier** that predicts the value of the (omitted) sensitive feature for the given CF sample. If the CF sample is classified as e.g. male (privileged group), while the original sample was e.g. female (unprivileged group), the decision model could be biased and its unfairness can be quantified (Eq. 3 and 4).

Indeed, each CF sample derives from the original sample $\mathbf{x}$ plus a perturbation ϵ, where ϵ is the *distance* from the original sample for getting a positive outcome, and it should be independent from the user-sensitive characteristics. Figure 1 depicts a scenario in which *male* (blu color) is the privileged category, and *female* (red color) is the unprivileged one. For each subfigure, a sample with an unfavorable decision and its corresponding CFs are depicted. A classic ML model (i.e., XGB) is compared with a debiasing ML model (i.e., AdvDeb). We can observe that for the male sample and classic ML model (Figure 1(a)), the CF samples belong to the same sensitive category (i.e., male). For the female sample (Figure 1 (b)), this is not true, revealing a bias of the model. Conversely, the debiasing model (Figure 1 (c) and (d)) shows no predominance in the generated counterfactuals of one value of the sensitive class. However, a change of the outcome, e.g. from negative to positive, should not be determined by a flip of the value(s) of the sensitive feature(s). Now, we introduce our fairness criteria and metrics.

Definition 3.1 (Counterfactual Fair Opportunity)**.** *A decision model is fair if the counterfactual samples of individuals with unfavorable decisions maintain the same sensitive value to reach a positive outcome. This behavior must be guaranteed both for the privileged and the unprivileged group [20].*

$$\mathbb{P}(f_s(\mathscr{C}_{\mathscr{X}|_{s=0}^-}) \neq s \mid f(\mathscr{C}_{\mathscr{X}|_{s=0}^-}) = 1, \mathscr{X}|_{s=0}^-) = \mathbb{P}(f_s(\mathscr{C}_{\mathscr{X}|_{s=1}^-}) \neq s \mid f(\mathscr{C}_{\mathscr{X}|_{s=1}^-}) = 1, \mathscr{X}|_{s=1}^-) \tag{1}$$

To define a sort of discrimination score of a given decision model, we propose a metric that we call *Counterfactual Flips*. The metric quantifies the discriminatory behavior the model might put in place.

Definition 3.2 (Counterfactual Flips)**.** *Given a sample* $\mathbf{x}$ *belonging to a demographic group* s *whose model output is denoted as* $f(\mathbf{x})$*, a generated set* $\mathscr{C}_\mathbf{x}$ *of* k *counterfactuals with desired* y^* *outcome.* $\forall \mathbf{c}_\mathbf{x}^i \in \mathscr{C}_\mathbf{x}$ *s.t.* $f(\mathbf{c}_\mathbf{x}^i) = y^*$*, the Counterfactual Flips indicate the percentage of counterfactual samples belonging to another demographic group (i.e.,* $f_s(\mathbf{c}_\mathbf{x}^i) \neq f_s(\mathbf{x})$*, with* $f_s(\mathbf{x}) = s$*).*

$$\text{CFlips}(\mathbf{x}, \mathscr{C}_\mathbf{x}, f_s(\cdot)) \triangleq \frac{\sum_{i=1}^k (\mathbb{1}(\mathbf{c}_\mathbf{x}^i))}{k} \quad \textit{where } \mathbb{1}(\mathbf{c}_\mathbf{x}^i) = \begin{cases} 1 & \textit{if } f_s(\mathbf{c}_\mathbf{x}^i) \neq f_s(\mathbf{x}) \neq s \\ 0 & \textit{if } f_s(\mathbf{c}_\mathbf{x}^i) = f_s(\mathbf{x}) = s \end{cases} \tag{2}$$

The bigger the CFlips value is, the stronger the bias the model suffers from. In our work, we only take into account samples negatively predicted by the decision maker (i.e., $f(\mathbf{x}) = 0$) as we are interested in quantifying the discrimination in achieving a positive counterfactual result (i.e., $f(\mathbf{c}_\mathbf{x}) = 1 \wedge f_s(\mathbf{c}_\mathbf{x}) \neq s$). Given a set of samples $\mathscr{X}^- \subseteq \mathscr{D}$ predicted by the decision maker as negative (unfavorable decision), the metric in Eq. 2 can be generalized to the *unprivileged* and *privileged* group (in Eq. 3 $s = 0$ for the *unprivileged* samples negatively predicted, and $s = 1$ for the *privileged* samples negatively predicted).

$$\text{CFlips}_s \triangleq \frac{\sum_{i=1}^n \text{CFlips}(\mathbf{x}_i, \mathscr{C}_{\mathbf{x}_i}, f_s(\cdot))}{|\mathscr{X}|_s^-|} \quad \text{with } \mathbf{x}_i \in \mathscr{X}|_s^- \tag{3}$$

A limitation of the CFlips metric is that it does not measure the distance of each CF sample from the original data point. However, from an individual-fairness wise, a debated issue is the definition of a metric that considers that distance [21]. Accordingly, we propose a new metric that considers CFs ranked based on the Mean Absolute Deviation from the original sample and other criteria [6]. The insight behind this metric is that the more the CF

is ranked high (in the top positions of the ranking), the more its impact on the metric value. Thus, the metric penalizes CFs ranked in the top positions for which the value of the sensitive feature is flipped. More formally:

Definition 3.3 (Discounted Cumulative Counterfactual Fairness)**.** Given a set of Counterfactuals $\mathbf{C_x}$ for a sample $\mathbf{x}_i$, the *Discounted Cumulative Counterfactual Fairness* $\text{DCCF}_{\mathbf{x}_i}$ measures the cumulative gain of the ranking of counterfactuals w.r.t. the sensitive group of the original sample:

$$\text{DCCF}_{\mathbf{x}_i} \triangleq \sum_{p_j, \mathbf{c}^j_{\mathbf{x}_i} \in \mathscr{C}_{\mathbf{x}_i}} \frac{2^{(1-\mathbb{1}(c^j_{\mathbf{x}_i}))} - 1}{\log_2(p_j + 1)} \tag{4}$$

where p_j is the rank of $\mathbf{c}^j_{\mathbf{x}_i}$ in $\mathscr{C}_{\mathbf{x}_i}$ and $\mathbb{1}(c^j_{\mathbf{x}_i})$ from Eq. 2.

If more CF samples belonging to the same sensitive group as the original data point are in a higher ranking position, the result will be a higher DCCF. Thereby, we can formulate the *Ideal Discounted Cumulative Counterfactual Fairness* (IDCCF) as an ideal ranking in which each CF sample $\mathbf{c_x}$ belongs to the same sensitive group as the original sample $\mathbf{x}$ (Eq. 5), and the *normalized* DCCF (nDCCF) (Eq. 6).

$$\text{IDCCF}_{\mathbf{x}_i} \triangleq \sum_{p_j, \mathbf{c}^j_{\mathbf{x}_i} \in \mathscr{C}_{\mathbf{x}_i}} \frac{2^{(1)} - 1}{\log_2(p_j + 1)} \tag{5}$$

$$\text{nDCCF}_{\mathbf{x}_i} \triangleq \frac{\text{DCCF}_{\mathbf{x}_i}}{\text{IDCCF}_{\mathbf{x}_i}} \tag{6}$$

In the same way as CFlips, given a set of samples $\mathscr{X}^- \subseteq \mathscr{D}$ predicted by the decision model as negative, the metric in Eq. 6 can be generalized to the *unprivileged* and *privileged* group (Eq. 7).

$$\text{nDCCF}_s \triangleq \frac{1}{|\mathscr{X}|^-_s|} \sum_{\mathbf{x}_i} \text{nDCCF}_{\mathbf{x}_i} \quad \text{with } \mathbf{x}_i \in \mathscr{X}|^-_s \tag{7}$$

For both CFlips and nDCCF, we are interested in the difference (i.e., Δ), between *privileged* and *unprivileged*, being close to zero.

3.2. Explainability through the counterfactual lens

Several methods have been proposed to explain black-box models. SHAP is inspired by the cooperative game theory based on the Shapley Values [22]. Each feature is considered a player that contributes differently to the outcome (i.e., the algorithm decision). However, the explanation provided by this method probably is not so clear for a customer who does not have experience with how an algorithm works. Furthermore, Shapley value does not give in to which extent changing a feature can result in a different outcome. For this reason, if we want to improve the user's trust and, in general, the user experience with the system, we need to make the explanation more understandable. Counterfactual Reasoning can be useful in that direction. Indeed A counterfactual $\mathbf{c_x}$ can be seen as a perturbation from a starting sample $\mathbf{x}$ of a quantity ϵ (i.e., $\mathbf{c_x} = \mathbf{x} + \epsilon$). For a numerical or ordinal feature i, ϵ_i can be expressed as the difference between the counterfactual and the feature of the sample $c_{x_i} - x_i$. For a categorical feature j, ϵ_j can be expressed in a *one-hot encoding* form as -1 to the category that is removed and 1 to the category that is engaged. Let be δ the difference between the posterior conditional probability of predicting a counterfactual sample and the original sample as belonging to the privileged group (i.e., $\delta = \mathbb{P}(f(\mathbf{c_x}) = 1|\mathbf{c_x}) - \mathbb{P}(f(\mathbf{x}) = 1|\mathbf{x})$). We can identify the most influential features for $f(\cdot)$ evaluating the Pearson correlation between ϵ and δ: $\rho(\epsilon, \delta)$. In the same way, we can identify the proxy feature influencing a discrimination in the decision maker through the investigation of $f_s(\cdot)$ [17, 23, 24]. The ranked correlation can be used to generate a Natural Language based explanation for the knowledge expert and a user-based explanation using the features of the nearest counterfactual sample (i.e., through the investigation of ϵ as actionable recommended step) [12, 25].

4. Experimental Analysis

4.1. Experimental setting

Dataset. The experimental evaluation has been carried out on state-of-the-art benchmark datasets (i.e., Adult[3] with *gender* as sensitive information). We do not include any sensitive features for training the model, guaranteeing the *fairness under unawareness* setting.

Decision Maker. To keep the approach as general as possible, we opted for Logistic Regression[4] (LR), Support-Vector Machines[4] (SVM), XGBOOST[4] (XGB) , and LightGBM[4] (LGBM).

Debiased Decision Maker. To investigate the quality and the reliability of our metrics we used also two debiased classifiers, *Adversarial Debiasing*[4] (AdvDeb) proposed by Zhang et al. [26] and *Linear Fair Empirical Risk Minimization*[4] (lferm) proposed by Donini et al. [27] as in-processing algorithms.

Counterfactual Generator. For the sake of reproducibility and reliability, the counterfactuals are generated with an external counterfactual framework, DiCE [6], with $|\mathscr{C}_{\mathbf{x}}|$ equal to 100[5].

Sensitive-Feature Classifier. We used XGB for implementing this component due to its capability to learn

[3] https://archive.ics.uci.edu/ml/datasets/adult

[4] LR, SVM: https://scikit-learn.org/; XGB: https://github.com/dmlc/xgboost; LGBM: https://github.com/microsoft/LightGBM; AdvDeb: https://github.com/Trusted-AI/AIF360; lferm: https://github.com/jmikko/fair_ERM;

[5] DiCE offers several strategies for generating candidate counterfactual samples, but we choose to only exploit the Genetic one.

Table 1
Accuracy, DEO, ΔCFlips (%), and ΔnDCCF metrics with XGB as $f_s(\cdot)$ on the Adult, Adult-debiased, Crime, and German test set. We mark the best-performing method for each metric in bold font.

	Adult (*gender*)					
			ΔCFlips ↓		ΔnDCCF ↓	
$f(\cdot)$	ACC↑	DEO↓	Genetic	KDtree	Genetic	KDtree
LR	0.8099	0.0546	67.05	76.03	0.6363	0.7476
DT	0.8161	0.0760	70.23	77.14	0.6303	0.7500
SVM	0.8541	0.0644	73.99	79.35	0.7223	0.7767
LGBM	0.8658	0.0379	70.87	78.40	0.6777	0.7723
XGB	**0.8698**	0.0635	70.40	78.42	0.6716	0.7717
RF	0.8534	0.0216	73.09	76.68	0.7017	0.7528
MLP	0.8494	0.0529	71.32	77.95	0.6862	0.7663
LFERM	0.8428	**0.0194**	32.40	68.25	0.2613	0.6434
ADV	0.8512	0.1399	**7.96**	53.90	**0.0725**	0.4870
FairC	0.8395	0.2451	38.72	**36.27**	0.3196	**0.2970**

non-linear dependencies.

Metrics. We evaluate the models' performance with the Accuracy (ACC) and model fairness by measuring Equal Opportunity[6] (DEO).

Split and Hyperparameter Tuning. The datasets have been split with the hold-out method 90/10 train-test set, with stratified sampling w.r.t. the target and sensitive labels, to respect the original distribution in each split. The Decision Maker, the Debiased models, and the Sensitive-Feature Classifier have been tuned on the training set with a Grid Search k-fold (k=5) cross-validation methodology, the first two optimizing AUC metric, and the latter F1 score to prevent unbalanced predictions on the sensitive feature.

4.2. Fairness Results

Now that the setting is clear enough, we can move on to analyze how well they perform in terms of fairness. The performance of the Decision Makers on the metrics DEO, as well as our suggested metrics CFlips and nDCCF are reported in Table 1. It is important to point out that the CFlips metric indicates how often a change of result for the Decision Maker corresponds to a change in the classification of the sensitive feature (e.g., from female to male and vice-versa). Conversely, the nDCCF metric gives more importance to counterfactuals with highest positions in the ranking (the most similar to the original sample) that do not change the sensitive class.

For the three debiased models (i.e., AdvDeb, lferm, and FairC) the Δ is close to zero for both our metrics, meaning that there is not a great difference in the CFlips for both groups (privileged and unprivileged one). The debiased models perform the same both with standard fairness metrics and our metrics (i.e., CFlips, nDCCF).

[6] $DEO = \left|\mathbb{P}(\hat{Y} = 1 \mid S = 1, Y = 1) - \mathbb{P}(\hat{Y} = 1 \mid S = 0, Y = 1)\right|$

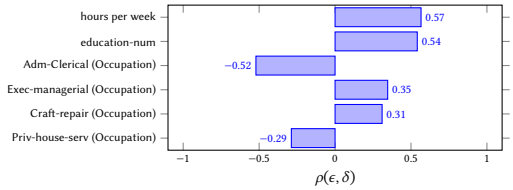


Figure 2: Top-6 most correlated features with a *gender* Flip (i.e., $\rho(\epsilon, \delta)$) on the Adult-debiased dataset with Genetic strategy as $g(\mathbf{x})$ and MLP as both $f(\cdot)$ and $f_s(\cdot)$.

4.3. Explainability Results

Following a brief analysis of how our methodology can be useful not only to investigate unfair model behaviour but also to explain and quantify proxy discriminative features.

In Figure 2, we can find the rank of features correlation with a Flip in $f_s(\cdot)$ with MLP as $f(\cdot)$ decision boundary for the generation of $\mathbf{c_x}$ and XGB as $f_s(\cdot)$ for the Adult-debiased dataset. The analysis is restricted to only samples negatively predicted in order to specifically quantify the *proxy-features* that lead to a positive prediction with also a change in the sensitive information. In detail, a negatively correlated feature (e.g., *Adm-Clerical*) is a feature that has an opposite direction with respect to $\mathbb{E}[f_s(\mathcal{X}^-) \mid \mathcal{X}^-]$ while a positively correlated one (e.g., *hours per week*) has the same direction.

5. Conclusion

In this work, we present a novel methodology for detecting bias in decision-making models that do not use sensitive features and work in a context of fairness under unawareness. Furthermore, we propose a new fairness concept (i.e., *Counterfactual Fair Opportunity*), two related fairness metrics (i.e., CFlis and nDCCF), and an explainability methodology.

In the future, we plan to define a strategy to generate fair and actionable counterfactual samples with the aim of developing a debiasing model that could be effectively fair in the context of *fairness under unawareness*.

References

[1] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, A. Huq, Algorithmic decision making and the cost of fairness, in: KDD, ACM, 2017, pp. 797–806.

[2] M. L. Ginsberg, Counterfactuals, Artif. Intell. 30 (1986) 35–79.

[3] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artif. Intell. 267 (2019) 1–38.

[4] J. Pearl, Causation, action and counterfactuals, in: ECAI, John Wiley and Sons, Chichester, 1994, pp. 826–828.

[5] R. Ferrario, Counterfactual reasoning, in: CONTEXT, volume 2116 of *Lecture Notes in Computer Science*, Springer, 2001, pp. 170–183.

[6] R. K. Mothilal, A. Sharma, C. Tan, Explaining machine learning classifiers through diverse counterfactual explanations, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 607–617.

[7] M. J. Kusner, J. R. Loftus, C. Russell, R. Silva, Counterfactual fairness, in: NIPS, 2017, pp. 4066–4076.

[8] J. Joo, K. Kärkkäinen, Gender slopes: Counterfactual fairness for computer vision models by attribute manipulation, in: Proceedings of the 2nd International Workshop on Fairness, Accountability, Transparency and Ethics in Multimedia, 2020, pp. 1–5.

[9] M. Tavakol, Fair classification with counterfactual learning, in: SIGIR, ACM, 2020, pp. 2073–2076.

[10] L. Bottou, J. Peters, J. Q. Candela, D. X. Charles, M. Chickering, E. Portugaly, D. Ray, P. Y. Simard, E. Snelson, Counterfactual reasoning and learning systems: the example of computational advertising, J. Mach. Learn. Res. 14 (2013) 3207–3260.

[11] A. Swaminathan, T. Joachims, Batch learning from logged bandit feedback through counterfactual risk minimization, J. Mach. Learn. Res. 16 (2015) 1731–1755.

[12] G. Cornacchia, F. Narducci, A. Ragone, A general model for fair and explainable recommendation in the loan domain (short paper), in: KaRS/ComplexRec@RecSys, volume 2960 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021.

[13] M. Dudík, J. Langford, L. Li, Doubly robust policy evaluation and learning, in: ICML, Omnipress, 2011, pp. 1097–1104.

[14] A.-H. Karimi, G. Barthe, B. Schölkopf, I. Valera, A survey of algorithmic recourse: Contrastive explanations and consequential recommendations 55 (2022). URL: https://doi.org/10.1145/3527848. doi:10.1145/3527848.

[15] A. Korikov, A. Shleyfman, J. C. Beck, Counterfactual explanations for optimization-based decisions fin the context of the GDPR, in: IJCAI, ijcai.org, 2021, pp. 4097–4103.

[16] K. Sokol, P. A. Flach, Counterfactual explanations of machine learning predictions: Opportunities and challenges for AI safety, in: SafeAI@AAAI, volume 2301 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019.

[17] G. Cornacchia, V. W. Anelli, G. M. Biancofiore, F. Narducci, C. Pomo, A. Ragone, E. Di Sciascio, Auditing fairness under unawareness through counterfactual reasoning, Information Processing & Management 60 (2023) 103224. URL: https://www.sciencedirect.com/science/article/pii/S0306457322003259. doi:https://doi.org/10.1016/j.ipm.2022.103224.

[18] J. Chen, N. Kallus, X. Mao, G. Svacha, M. Udell, Fairness under unawareness: Assessing disparity when protected class is unobserved, in: FAT, ACM, 2019, pp. 339–348.

[19] A. Fabris, A. Esuli, A. Moreo, F. Sebastiani, Measuring fairness under unawareness via quantification, CoRR abs/2109.08549 (2021). URL: https://arxiv.org/abs/2109.08549. arXiv:2109.08549.

[20] G. Cornacchia, V. W. Anelli, F. Narducci, A. Ragone, E. D. Sciascio, Counterfactual reasoning for decision model fairness assessment, in: Companion of The Web Conference 2023, Austin, TX, USA, April 30 - May 4, 2023, ACM, 2023. doi:https://doi.org/10.1145/3543873.3587354.

[21] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. S. Zemel, Fairness through awareness, in: ITCS, ACM, 2012, pp. 214–226.

[22] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 4768–4777.

[23] G. Cornacchia, V. W. Anelli, F. Narducci, A. Ragone, E. D. Sciascio, Counterfactual fair opportunity: Measuring decision model fairness with counterfactual reasoning, 2023. arXiv:2302.08158.

[24] G. Cornacchia, V. W. Anelli, F. Narducci, A. Ragone, E. D. Sciascio, Counterfactual reasoning for bias evaluation and detection in a fairness under unawareness setting, 2023. arXiv:2302.08204.

[25] G. Cornacchia, F. Narducci, A. Ragone, Improving the user experience and the trustworthiness of financial services, in: C. Ardito, R. Lanzilotti, A. Malizia, H. Petrie, A. Piccinno, G. Desolda, K. Inkpen (Eds.), Human-Computer Interaction – INTERACT 2021, Springer International Publishing, Cham, 2021, pp. 264–269.

[26] B. H. Zhang, B. Lemoine, M. Mitchell, Mitigating unwanted biases with adversarial learning, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 335–340. URL: https://doi.org/10.1145/3278721.3278779. doi:10.1145/3278721.3278779.

[27] M. Donini, L. Oneto, S. Ben-David, J. Shawe-Taylor, M. Pontil, Empirical risk minimization under fairness constraints, in: NeurIPS, 2018, pp. 2796–2806.